Reliable application of the MATH taxonomy sheds light on assessment practices

Teaching Mathematics and Its Applications (2019) Page 1 of 16
doi:10.1093/teamat/hrz017

Reliable application of the MATH taxonomy sheds light on assessment practices

GEORGE KINNEAR†, MAX BENNETT, RACHEL BINNIE, RÓISÍN BOLT AND YINGLAN ZHENG

School of Mathematics, University of Edinburgh

[Received on 17 July 2019; revised on 6 November 2019; accepted on 20 November 2019]

The MATH taxonomy classifies questions according to the mathematical skills required to answer them. It was created to aid the development of more balanced assessments in undergraduate mathematics, and has since been used to compare different assessment regimes across school and university. To date, there has been no systematic investigation of the reliability of the taxonomy when applied by multiple coders, and it has only been applied in a limited range of contexts. In this paper we outline a calibration process which enabled four novice coders to attain a high level of inter-rater reliability. In addition, we report on the results of applying the taxonomy to different secondary school exams and to all assessment questions in a first-year university mathematics module. The results confirm previous findings that there is a difference between the mix of skills assessed in school and university mathematics exams, although we find a notably different assessment profile in the university module than in previous work. The calibration process we describe has the potential to be used more widely, enabling reliable use of the MATH taxonomy to give insight into assessment practices.

Keywords: MATH taxonomy; assessment; constructive alignment

1. Introduction

Exams are important – not only do they provide credentials that certify the examinees' abilities, they have a 'backwash' effect on teaching and learning (Gibbs & Simpson, 2004). In university mathematics, the closed-book exam is the most common form of assessment, with a recent survey of modules offered at UK institutions finding "nearly 70% ... use closed book examinations for at least three quarters of the final mark" (Iannone & Simpson, 2012, p4). It is important to understand what skills are being assessed in these exams, in order to judge how well they are aligned with the desired learning outcomes (Biggs & Tang, 2011). Similarly, there are concerns about the standards in school mathematics exams, particularly regarding their role as preparation for further study of mathematics (Darlington, 2015b, Section 1.1).

Question taxonomies can provide useful insight into the balance of skills being assessed in an exam. We give an overview of different taxonomies in §2.1, but among mathematics-specific taxonomies, the "MATH taxonomy" (Smith *et al.*, 1996) is the most widely-used. In particular, it has been used to compare the skills assessed in a variety of exams at secondary and tertiary level (Darlington, 2014, 2015a,b).

Here, we report on an undergraduate project in which four mathematics students (the final four authors) developed a shared understanding of the MATH taxonomy and applied it to a wide range of exams. This is the first work to explicitly investigate inter-rater reliability in the application of the MATH taxonomy. It also extends the work of Darlington (2014; 2015a) by considering a larger sample

†Corresponding author. Email: G.Kinnear@ed.ac.uk

of exams, including a new university context. Furthermore, we use the MATH taxonomy to analyse and compare all the components of assessment within a university module. In summary, we address the following research questions:

1. To what extent can the MATH taxonomy be applied reliably by novice coders?
2. Are there differences between the skills assessed in
   (a) secondary and tertiary exams?
   (b) secondary exams from different education systems?
   (c) different components of a university module?

After a review of relevant literature in the next section, we describe the calibration process used by the four novice coders to achieve a high level of reliability. Finally, we present the results of independent coding of a set of 58 different exams (drawn from A-Level, IB, SQA Advanced Higher and the University of Edinburgh) and discuss the implications.

2. Literature review

2.1 *Taxonomies*

Various taxonomies have been developed for classifying educational tasks, with perhaps the most well-known being Bloom's taxonomy (Bloom *et al.*, 1956) and the SOLO taxonomy (Biggs & Collis, 1982). In practice, these have been found difficult to apply to mathematics (see Darlington (2015a, Section 2.2) for a discussion), leading to the development of mathematics-specific taxonomies. The Mathematical Assessment Task Hierarchy, also known as the "MATH taxonomy", was introduced to help teachers of mathematics in higher education to construct more balanced assessments (Smith *et al.*, 1996). The MATH taxonomy classifies tasks according to the skills needed to solve them, with various categories arranged in three broad groups; these are summarised in Table 1. A similar classification scheme was later developed independently by Pointon & Sangwin (2003) and by Tallman *et al.* (2016). The similarity of these taxonomies supports the validity of the approach.

In a similar way, Lithner (2008) classifies tasks according to different reasoning types, e.g. "algorithmic reasoning" and "creative mathematically founded reasoning". This framework has been used to classify tasks in university exams (Bergqvist, 2007) and upper secondary school exams (Boesen *et al.*, 2010) in Sweden.

2.2 *Applications of the MATH taxonomy*

The MATH taxonomy has been applied to university courses, with early findings suggesting that most exam papers are "heavily biased towards Group A tasks" (Ball *et al.*, 1998, p828). However, it is likely that the intended learning outcomes from these courses will include Group C skills. According to constructive alignment (Biggs & Tang, 2011), a widely-used theoretical basis for designing teaching in higher education, the learning outcomes from the course should be addressed directly through the teaching and assessment. The MATH taxonomy is positioned as a helpful tool to enable this:

> "The MATH taxonomy provides a convenient reference table for checking that the students' experiences are suitably diversified. A survey of most tests and examinations reveals a preponderance of examples testing a narrow range of student performance." (Ball *et al.*, 1998, p840)

TABLE 1. *Summary of the MATH taxonomy, adapted from Darlington (2015a)*

Group	Outline	Subgroup	
A	Factual recall and routine procedures	FKFS	Factual knowledge and fact systems
		COMP	Comprehension
		RUOP	Routine use of procedures
B	Using existing mathematical knowledge and techniques in new ways	IT	Information transfer
		AINS	Application in new situations
C	Application of conceptual knowledge to construct mathematical arguments	JI	Justifying and interpreting
		ICC	Implications, conjectures and comparisons
		EV	Evaluation

In a study of 52 exam papers from year 1 courses at Russell Group universities, Darlington (2015b) found that "The majority of marks awarded in undergraduate papers were for Group C skills (51.7%); however, there were also 44.1% for Group A skills, the vast majority of which were for factual recall" (p190). This motivates RQ2(c) and our in-depth study of one university module (reported in §4.4), using the MATH taxonomy to understand the extent to which the course assesses a diverse range of skills.

The MATH taxonomy has also been used to study the school-university interface. In particular, Ellie Darlington has applied the MATH taxonomy to school and university exams (Darlington, 2014, 2015a), and to university admissions tests (Darlington, 2015b). This work has established a consistent pattern of school-level exams focusing heavily on Group A skills; for instance "The vast majority (89.9%) of marks awarded in A-level examinations were for Group A skills" (Darlington, 2015b, p188). This motivates RQ2(a) and RQ2(b), in that we wish to confirm this pattern across a wider range of contexts (as described in §3.1).

One aspect of the MATH taxonomy which has not been well-explored is its connection with student performance. In introducing the MATH taxonomy, Smith *et al.* (1996) stressed that "no hierarchy of difficulty is implied as we move down the list" (p68), however Wood & Smith (2002) found that when students were asked to rank eight tasks (one from each MATH category) in order of perceived difficulty, their ranking corresponded closely with the order of the MATH taxonomy. On the other hand, Wood *et al.* (2002) found that student performance on questions from different MATH Groups was well-correlated: "the correlation between A% and the average of B and C% is a very high 0.83" (p6).

TABLE 2. *Examples of questions classified in this study.*

Category	School example	University example
	Group A	
FKFS	Write down the derivative of $\sin^{-1} x$.	*No examples observed*
COMP	Obtain the matrix, A, associated with an anti-clockwise rotation of $\frac{\pi}{3}$ radians about the origin.	Which of the following matrices are orthogonal? (a) $\begin{pmatrix} 4 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$, (b) $\begin{pmatrix} \cos\left(\frac{\pi}{7}\right) & -\sin\left(\frac{\pi}{7}\right) \\ \sin\left(\frac{\pi}{7}\right) & \cos\left(\frac{\pi}{7}\right) \end{pmatrix}$, ...
RUOP	Use partial fractions to find $\int \frac{3x-7}{x^2-2x-15}\,dx$.	Find the eigenvalues and corresponding eigenvectors of the matrix $\begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix}$.
	Group B	
IT	Given $z = x+iy$, sketch the locus in the complex plane given by $\lvert z\rvert = \lvert z-2+2i\rvert$.	For what value of k do the following vectors NOT form a basis of $\mathbb{R}^3$? $\begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 16 \\ -14 \\ k \end{pmatrix}, \begin{pmatrix} -3 \\ 3 \\ -2 \end{pmatrix}$.
AINS	Prove directly that the sum of any three consecutive integers is divisible by 3.	Given that $\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \cdot A \cdot \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} = 5 \cdot I$, evaluate the determinant of matrix A.
	Group C	
JI	Prove by induction that $\sum_{r=1}^{n} r!r = (n+1)!-1$ for all positive integers n.	"Differentiation of polynomials can be regarded as a linear map". Explain briefly with an appropriate example why this is so.
ICC	Let n be a positive integer. Find a counterexample to show that the following statement is false: "n^2+n+1 is always a prime number".	Give an example of a three by three matrix with real entries (i.e. in $\mathbb{R}^{3\times 3}$), with exactly one real eigenvalue which is non-zero, and also with two complex eigenvalues. Repeated eigenvalues will be counted with multiplicity.
EVAL	The compactness, C , of an enclosed region can be defined by $C = \frac{4A}{\pi d^2}$, where A is the area of the region and d is the maximum distance between any two points in the region. [...] Comment briefly on whether C is a good measure of compactness.	*No examples observed*

2.3 *Inter-rater reliability*

Classifying tasks using the MATH taxonomy requires judgement, and in order to draw robust conclusions from the resulting classifications we should check that these judgements are being performed reliably by multiple coders. This issue is not particular to the MATH taxonomy; for instance a recent survey (Coleman, 2017) investigated the use of inter-rater reliability statistics in studies making use of educational taxonomies (predominantly Bloom's taxonomy). The survey concluded that "In order to prove that research using other educational taxonomies can provide a sound evidence base for qualifications evaluation, comparability and development, further targeted studies will be necessary" (p36).

There are numerous measures of reliability in use, as can be seen in the table provided by Coleman (2017, p31) or the discussion in Hayes & Krippendorff (2007). Here, we make use of Krippendorff's α since this measure "can handle multiple coders; nominal, ordinal, interval, ratio, and other metrics; missing data; and small sample sizes" (Krippendorff, 2004, p428) and thus has the potential to provide a standard measure of reliability across a range of applications. Perfect agreement is represented by the maximum value of $\alpha = 1$, while perfect disagreement corresponds to $\alpha = 0$. The standard advice for interpreting intermediate values is:

> "To assure that the data under consideration are at least similarly interpretable by two or more scholars (as represented by different coders), it is customary to require $\alpha \geqslant .800$. Where tentative conclusions are still acceptable, $\alpha \geqslant .667$ is the lowest conceivable limit" (Krippendorff, 2004, p429)

3. Method

3.1 *Materials*

In order to address RQ2, we selected a range of school and university assessment questions to be coded. A summary of the exam papers used is given in Table 3.

For school exams, we chose to study exams which are common entry qualifications at the University of Edinburgh: A-Level Mathematics, International Baccalaureate Higher Level Mathematics (IB HL), and SQA Advanced Higher Mathematics (SQA AH). Within A-Level Mathematics, students take exams in a number of modules and from these we chose to study modules C1-C4 and FP3; this is in line with Darlington (2015a) (which considers C1 and FP3) but extends the range of Core modules beyond C1. The A-Level exams are offered by a number of different exam boards; we chose the largest of these (Edexcel) which is taken by a majority of the cohort (54.8% in 2017/18, (Ofqual, 2019, p12)).

At the post-secondary level, we studied one university course in depth: Introduction to Linear Algebra (ILA), which is a year 1, 20-credit course in a year of 120 credits (for some further details, see (Sangwin, 2018)). This was chosen as it is the first course in the mathematics programme at the University of Edinburgh, similar to Darlington (2014) which considers the University of Oxford. The MATH taxonomy was applied to the exam questions from this course (to address RQ2(a)), and also to other assessment questions (to address RQ2(c)). A final open-book exam contributes 80% to students' grade for the course, with the remainder made up of a variety of online and written assessments. In addition to this course, we also considered both a university admissions test, the Test of Mathematics for University Admissions (TMUA) (Gilbey & Robson, 2018), and the Diagnostic Test which is given to all students enrolled on year 1 mathematics modules at the University of Edinburgh (Kinnear, 2018).

TABLE 3. *Summary of the papers that were analysed, showing the total number of questions (meaning parts of questions to which marks were assigned in the paper) and marks.*

Exam	Papers		Total questions	Total marks
A-Level (Edexcel)	C1, 2014-2017	4	93	300
	C2, 2014-2017	4	95	300
	C3, 2014-2017	4	95	300
	C4, 2014-2017	4	88	300
	FP3, 2014-2017	4	78	300
SQA AH	2012-2019	8	244	800
IB HL	2014-2017 (P1 & P2)	8	232	920
TMUA	2016-2018 (P1 & P2)	6	120	120
DiagTest	-	1	24	100
ILA	2011-2018 (Dec/Aug)	13	393	1760
		58	1462	5200

3.2 *Coding procedure*

Starting with no knowledge of the MATH taxonomy, the four coders first read and discussed its definition (Smith *et al.*, 1996) and the examples provided by Darlington (2014). A single category was chosen for each part of a question that was assigned a specific mark in the paper. This meant that some questions were classified in Groups B and C even though some of the necessary working will have been routine calculations. The context of a question was also considered: as noted by Darlington (2014, Section 2.3.3), the skill being tested may be different depending on when the question is asked in a course, or the particular background of the student. The four coders were familiar with the contexts considered in this study, and used this experience to inform their judgment about how a typical student would be likely to respond.

In order to establish a common understanding of the MATH taxonomy, and thereby achieve reliable coding, an iterative process was used. This consisted of rounds of semi-independent coding followed by group discussion, using a series of different sets of questions, as summarised in Figure 1. These phases were all followed by a group discussion of any disagreements in order to arrive at a consensus and to clarify the coders' shared understanding of the taxonomy. To provide information about how reliably the coders were operating, Krippendorff's α was also computed after each phase. This was calculated using the `kripp.alpha` function from the R package `irr` (Gamer *et al.*, 2019), with the MATH categories treated as nominal-level data.

In Phases 1 and 2, the coders worked semi-independently, meaning that they were each tasked with producing their own coding of the questions but some discussion between coders was allowed to help establish a common understanding of the taxonomy. At Phase 3, the four coders were split into two pairs so that discussion could take place within the pair, but the two pairs worked independently to allow some indication of their emerging inter-rater reliability. At Phase 4, the pairs coded different sets of questions and the coders within the pairs worked independently; inter-rater reliability was checked for each pair.

At this point, coders had established a good working understanding of the taxonomy and high inter-rater reliability, so they proceeded to independent coding of different sets of questions from Table 3. As a final check of inter-rater reliability, all four coders independently coded a new exam paper. The process concluded with all coders reviewing the questions they had coded in earlier phases and updating

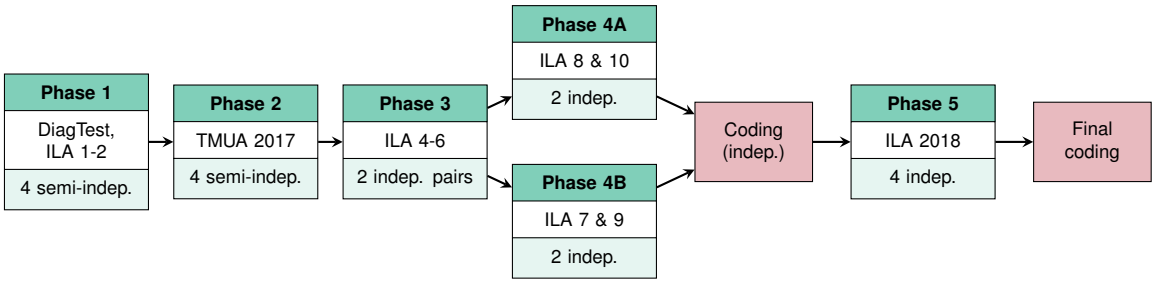

FIG. 1. Summary of the calibration and coding process. At each phase, a different set of course materials was coded by the coders, who mostly worked independently but in some cases (marked "semi-indep.") they conferred.

codes in line with established norms.

4. Results

4.1 *Inter-rater reliability*

The values of Krippendorff's α after each calibration phase are summarised in Table 4. Recalling that $\alpha \geqslant 0.8$ is generally regarded as a good level of inter-rater reliability, the results show that the coders attained a high level of inter-rater reliability over the course of the calibration process, reaching $\alpha > 0.9$ prior to embarking on independent coding. This high level of inter-rater reliability was sustained through to the final check (Phase 5) carried out after independent coding, when all four coders independently coded the same new exam with a high level of inter-rater reliability ($\alpha = 0.94$).

The bottom row of Table 4 shows the proportion of questions where the ultimate code assigned to the question (in the final coding phase, shown at the end of the process in Figure 1) was one that had been assigned by one of the coders in the calibration phase. This gives some measure of how stable the coding became – despite coders shifting their views and agreeing to go back and re-code certain types of question late in the process, this only affected a minority of previously coded items.

In the early phases, the most common disagreements arose from two issues. The first issue was the distinction between routine procedures (RUOP) and extrapolation of previously seen procedures to new situations (AINS). This depended on the coders' judgement of what could reasonably be expected to be "routine" for students, which developed through discussion and as more coding was carried out. The second issue was that both COMP and IT can include questions which involve checking whether conditions of a definition are satisfied, based on whether the definition is "simple" (COMP) or "conceptual"

TABLE 4. *Results at each phase of calibration, showing the number of items coded, the value of Krippendorff's alpha, and the percentage of items where one of the original codes was retained in the final coding phase.*

Phase	1	2	3	4A	4B	5
Items	32	40	43	18	27	21
α	0.74	0.73	0.93	0.92	0.95	0.94
% retained	75%	88%	58%	89%	81%	100%

(IT). The original description of the MATH taxonomy notes that that this distinction "will often be a subjective judgement" (Smith *et al.*, 1996, p69). The coders resolved this by discussing the definitions that appeared in the courses and developing a shared categorisation of these as simple or conceptual, which was consulted in subsequent coding.

4.2 *Comparing secondary and tertiary exams*

4.2.1 *Marks by Group.* The proportion of marks available in Groups A, B and C were calculated for all 58 exam papers; these are summarised by the mean values for each type of exam in Figure 2, with details for each paper given in the Appendix. This shows that school exams tend to have a higher proportion of marks available for Group A skills. A Bayesian version of a t-test (Kruschke, 2013) confirms this difference – the mean proportion of Group A marks for school exams is 0.68 (95% HDI [0.64, 0.72]) while for ILA[1] it is 0.56 (95% HDI [0.51, 0.60]). The difference between these means is 0.12 (95% HDI [0.06, 0.18]), and the fact that the 95% HDI is strictly positive confirms that there is a credible difference between the mean proportion of Group A marks in school exams and in ILA.

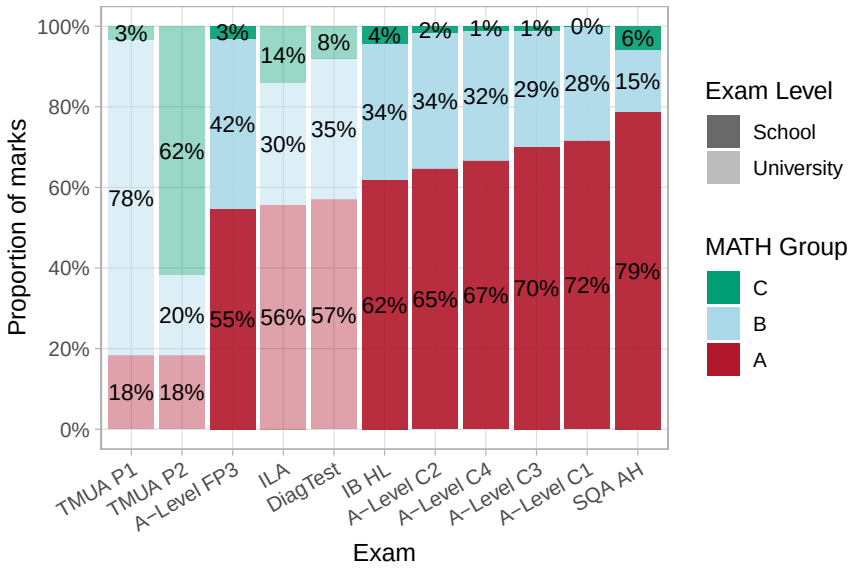

FIG. 2. Mean percentage of marks in each exam classified by MATH Group. School exams are shown with darker shading than university/admissions exams. The exams are ordered by mean proportion of Group A marks, with the TMUA paper having the least Group A marks, and SQA AH having the most.

4.2.2 *Marks by Category.* Looking at the more fine-grained MATH Categories, we find there is little difference between secondary and tertiary exams in terms of the distribution of marks within the categories, as shown in Table 5. Looking at Group A in particular, there is a notable difference with the

[1]We do not include TMUA and DiagTest since they are not exams taken for credit in higher education; in any case, adding them to the analysis does not affect the result.

findings of Darlington (2014), where 90% of the Group A marks were FKFS in the undergraduate exam. As shown in Figure 3, the university exam ILA has none of the available marks in the FKFS category. This is likely due to the fact that the ILA exam is open-book (with students permitted to bring a copy of the textbook and other notes into the exam), so the exam does not include any questions testing students' recall of definitions or theorems.

TABLE 5. *Proportion of available marks in each MATH Group by Category.*

Group	A			B		C		
Category	FKFS	COMP	RUOP	IT	AINS	JI	ICC	EVAL
Marks within group (%)								
A-Level C1	0	12	88	48	52	-	-	-
A-Level C2	0	11	89	30	70	100	0	0
A-Level C3	0	14	86	31	69	67	33	0
A-Level C4	0	13	87	19	81	100	0	0
A-Level FP3	0	3	97	11	89	100	0	0
SQA AH	0	3	97	61	39	91	9	0
IB HL	0	4	96	50	50	97	0	3
TMUA P1	0	0	100	43	57	50	50	0
TMUA P2	0	9	91	58	42	73	27	0
DiagTest	0	5	95	86	14	0	100	0
ILA	0	9	91	45	55	40	60	0

4.3 *Comparing secondary exams*

We now focus on the comparison between A-Level, SQA AH and IB HL papers. For this purpose, we group together all A-Level exams into one category; we believe this is justified as a typical student entering the University of Edinburgh will have studied all of these modules (with the possible exception of FP3) making this more directly comparable with SQA AH and IB HL. Figure 4 shows the distribution of the proportion of marks in each paper that were coded in each MATH Group. From this we see that even within one qualification, the various papers have a wide range of proportions of Group A marks. The other striking feature is that the mean proportion for SQA AH (0.79) appears to be much higher than both IB HL (0.62) and A-Level (0.66). This impression is supported by a Bayesian t-test: the estimated differences in proportions are 0.13 (95% HDI [0.05, 0.21]) with A-Level and 0.17 (95% HDI [0.04, 0.29]) with IB HL, and since the 95% HDIs are strictly positive there is a credible difference in the proportions.

4.4 *Components of a university module*

The MATH taxonomy was applied to all assessment items from the year 1 undergraduate module, Introduction to Linear Algebra (ILA). Each week, students complete three types of assessment which contribute part of the grade for the course:

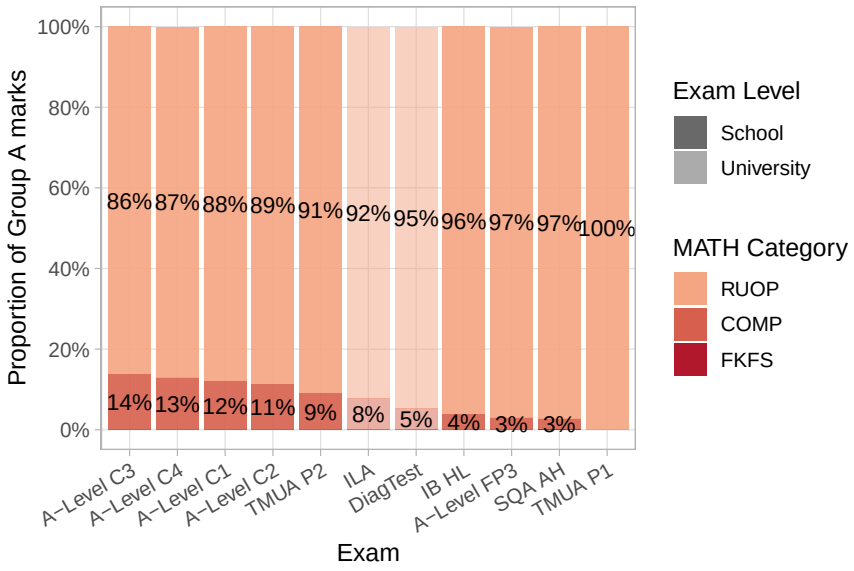

FIG. 3. Proportion of available marks in Group A falling in each MATH Category. School exams are shown with darker shading than university/admissions exams. The exams are ordered by mean proportion of RUOP marks. Note that no FKFS marks were observed.

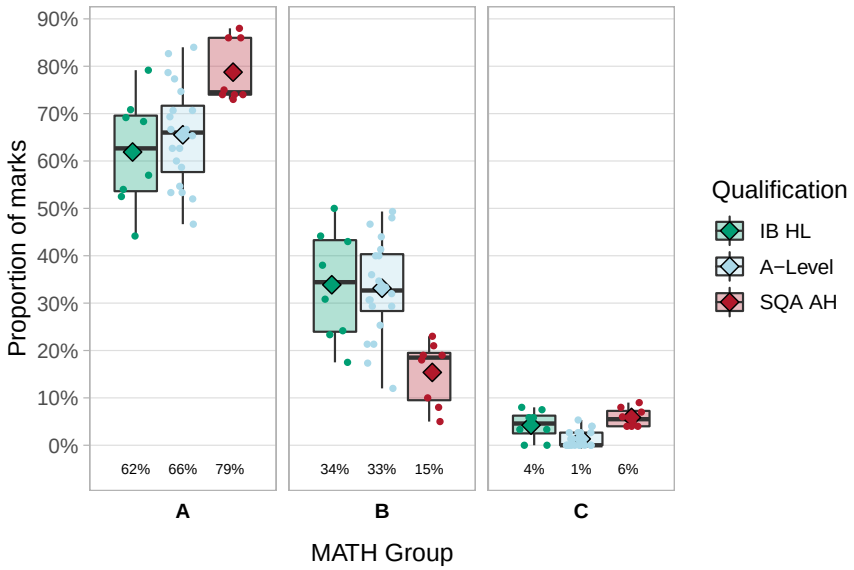

FIG. 4. Distribution of the percentage of marks in each MATH Group by Qualification. Each data point is an exam paper, with the distributions summarised by box plots and means (indicated by diamonds and with values shown below).

1. *reading quizzes*, which are short online tests about basic ideas from the week's reading,
2. *online assessments*, which consist of 4-7 questions on key procedures and concepts, and
3. *written assessments*, which are typically 2-3 longer questions requiring some sustained calculation and argument.

The first two components are delivered through the STACK computer-aided assessment system (Sangwin, 2013), which provides automatic marking and feedback. The written assessments are hand-marked by tutors and returned to students with feedback in weekly workshops.

Figure 5 shows the distribution of marks in each MATH Group for each component, as well as the distribution in the exam papers already considered in §4.2, while Table 6 gives the breakdown by MATH Categories. These show that the reading quizzes have the highest proportion of Group A tasks (77%) while the weekly written assessments have the lowest (28%). This is as expected, since the reading quizzes focus mainly on comprehension of new definitions (COMP accounts for 35% of the Group A marks) and practice with new procedures (RUOP is 65%). Assessment of these routine procedures is a strength of online assessment systems such as STACK, and enables the written assessments to make the most of tutor marking time by minimising the amount of routine work. The reading quizzes were 23% Group B, all of which were classified as IT as they were about "deciding whether or not conditions of a conceptual definition are satisfied" (Smith *et al.*, 1996, p69); an example is shown in Table 2.

In terms of Group C skills, the written assessments have a similar proportion of marks on these as the exam. Interestingly, the online assessments also have a nontrivial proportion of marks in Group C, the majority of which are due to questions on "the construction of examples and counterexamples" (Smith *et al.*, 1996, p70), such as the ICC example shown in Table 2. This is a strength of computer-aided assessment, with systems such as STACK able to check the mathematical properties of a student's answer using a computer algebra system (Sangwin, 2003).

TABLE 6. *Proportion of available marks in each MATH Group by Category, for components of assessment in ILA. The Coursework row is the (equally-weighted) mean of the three coursework components.*

Group	A			B		C		
Category	FKFS	COMP	RUOP	IT	AINS	JI	ICC	EVAL
Marks within group (%)								
Exam	0	9	91	45	55	40	60	0
Coursework	0	25	75	66	34	35	65	0
Reading	0	35	65	100	0	-	-	-
Online	0	19	81	83	17	20	80	0
Written	0	20	80	15	85	50	50	0

5. Discussion

We found that the MATH taxonomy can be applied reliably by novice coders, with an iterative process of calibration leading to four independent coders achieving high inter-rater reliability (Krippendorff's

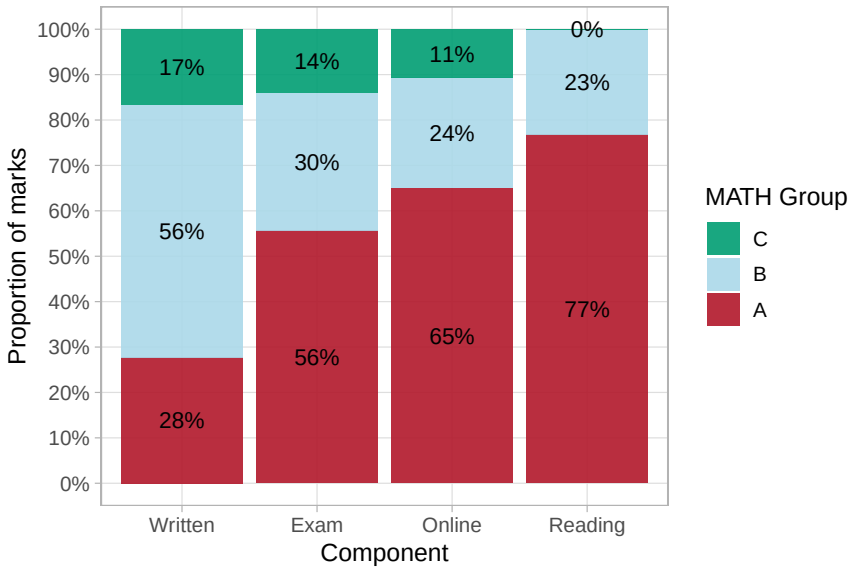

FIG. 5. Proportion of marks in each MATH Group for the four components of assessment in ILA.

$\alpha = 0.94$).

In applying the MATH taxonomy to a range of mathematics exams across secondary and tertiary levels, we found that there is a credible difference in the proportion of Group A skills assessed in school exams compared with a first-year university exam, and a credible difference in the mix of skills assessed in the three different school exams we considered (with a higher proportion of Group A marks available in SQA AH than in A-Level or IB HL). However, caution is needed in making a comparison between different school exams, as they serve slightly different purposes within their respective education systems (for instance, A-Level C1 mostly assesses content which is not in SQA AH because it is assessed in the prior qualification, SQA Higher).

Our application of the MATH taxonomy to all aspects of assessment in a particular university module showed that the balance of skills assessed in each component appears to be well-aligned with the lecturers' intentions. We also found that the balance of skills assessed in the exam appears to be quite different from that in many other university mathematics exams. In particular, Darlington (2015b) found that in a sample of 52 first-year undergraduate exams from 6 universities, the marks were "44.1% for Group A skills, the vast majority of which were for factual recall" (p190), while our results for ILA show that there were no factual recall marks. This discrepancy is likely due to the open-book rubric for the exam we considered, and in future work it would be interesting to examine a wider range of university exams.

5.1 *Limitations*

While we have good evidence that the raters developed a common understanding of the MATH taxonomy (from their high inter-rater reliability), there are two aspects of the validity of the coding which should be explored in future work. First, we cannot be sure that the codings will agree with other

applications of the taxonomy. Indeed, there are some differences between our results for A-Level Mathematics and those of Darlington (2015a), in that we found a much lower proportion of marks for Group A skills (66% compared with 90%). This could be partly caused by our decision to assign a single classification to each question part, rather than considering each individual mark on the marking scheme (since many questions were classified as Group B or C even though some of the marks were for routine calculation). The difference in Group A marks may also be due to the different exam boards considered in each case, as previous work found large differences between papers (Darlington, 2014, Table 3). Future work could explore a possible difference in the mix of skills assessed by different A-Level exam boards. The second aspect of validity is whether the codes assigned to questions actually match with the cognitive processes employed by students when solving them. Gierl (1997) investigated this issue for Bloom's taxonomy, and found that "The cognitive processes expected by item writers matched the processes used by students in only 54% of the cases" (p30).

A further possible concern is that raters were not blind to the source of the questions they were coding. This is an inherent difficulty with the MATH taxonomy – the context of the question is an important consideration as "one question may be considered routine in one particular instance, and yet non-routine in another" (Darlington, 2014, p215), so raters do need to understand the context in which a question is being asked (e.g. what students are expected to have learned so far, and whether certain types of questions appear predictably in exams). Indeed, the four coders here all had a good understanding of the context as they had recently studied these courses, which may have contributed to the high inter-rater reliability. Future work could investigate the extent to which coders are biased by the context, for instance by putting isolated questions from two different contexts (such as SQA AH and A-Level) into a common format, and presenting them to two independent coders assuming different contexts.

6. Conclusion

This work has demonstrated for the first time that the MATH taxonomy can be applied reliably by multiple coders. The iterative process used to develop common understanding among coders (described in §3.2) provides a model for establishing a group of reliable coders, which enables wider use of the MATH taxonomy in future. Our application of the MATH taxonomy to several secondary and tertiary exams has confirmed previous findings that school exams tend to assess more routine (Group A) skills than university exams, and our analysis of the different assessment components in a university module has shed light on how each component contributes to the constructive alignment of the course. There is scope for further work to explore the relationship between MATH categories and student performance, similar to Wood *et al.* (2002) and Gierl (1997), and to expand the use of the MATH taxonomy as a tool for understanding the mix of skills assessed in different university modules.

References

BALL, G., STEPHENSON, B., SMITH, G., WOOD, L., COUPLAND, M. & CRAWFORD, K. (1998) Creating a diversity of mathematical experiences for tertiary students. *International Journal of Mathematical Education in Science and Technology*, **29**, 827–841.

BERGQVIST, E. (2007) Types of reasoning required in university exams in mathematics. *The Journal of Mathematical Behavior*, **26**, 348–370.

BIGGS, J. B. & COLLIS, K. F. (1982) *Evaluating the Quality of Learning: The SOLO Taxonomy*. *Evaluating the Quality of Learning: The SOLO Taxonomy*.

BIGGS, J. & TANG, C. (2011) *Teaching for quality learning at university: what the student does*. McGraw-Hill/Society for Research into Higher Education/Open University Press.

BLOOM, B. S., ENGLEHART, M., FURST, E. J., HILL, W. H. & KRATHWOHL, D. R. (1956) *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. Longman.

BOESEN, J., LITHNER, J. & PALM, T. (2010) The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics*, **75**, 89–105.

COLEMAN, V. (2017) On the reliability of applying educational taxonomies. *Research Matters: A Cambridge Assessment publication*, 30–37.

DARLINGTON, E. (2014) Contrasts in mathematical challenges in A-level Mathematics and Further Mathematics, and undergraduate mathematics examinations. *Teaching Mathematics and its Applications*, **33**, 213–229.

DARLINGTON, E. (2015a) Post-16 Mathematics qualifications: Differences between GCE A level, International A level, Cambridge Pre-U and Scottish examination questions. *Research Matters: A Cambridge Assessment publication*, 6–13.

DARLINGTON, E. (2015b) What benefits could extension papers and admissions tests have for university mathematics applicants? *Teaching Mathematics and its Applications*, **34**, 179–193.

GAMER, M., LEMON, J., FELLOWS, I. & SINGH, P. (2019) *irr: Various Coefficients of Interrater Reliability and Agreement*. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

GIBBS, G. & SIMPSON, C. (2004) Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education*, **1**, 3–31.

GIERL, M. J. (1997) Comparing Cognitive Representations of Test Developers and Students on a Mathematics Test With Bloom's Taxonomy. *The Journal of Educational Research*, **91**, 26–32.

GILBEY, J. & ROBSON, D. (2018) The PRAC Taxonomy for Formal Mathematical Assessments. *Informal Proceedings of the 9th British Congress of Mathematics Education 2018* (S. Pope ed.). University of Warwick, pp. 55–58.

HAYES, A. F. & KRIPPENDORFF, K. (2007) Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, **1**, 77–89.

IANNONE, P. & SIMPSON, A. (2012) A Survey of Current Assessment Practices. *Mapping University Mathematics Assessment Practices* (P. Iannone & A. Simpson eds). University of East Anglia, pp. 3–15.

KINNEAR, G. (2018) Improving an online diagnostic test via item analysis. *Proceedings of the Fifth ERME Topic Conference on Mathematics Education in the Digital Age*. University of Copenhagen, pp. 315–316.

KRIPPENDORFF, K. (2004) Reliability in Content Analysis. *Human Communication Research*, **30**, 411–433.

KRUSCHKE, J. K. (2013) Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, **142**, 573–603.

LITHNER, J. (2008) A research framework for creative and imitative reasoning. *Educational Studies in Mathematics*, **67**, 255–276.

OFQUAL (2019) Annual qualifications market report: academic year 2017 to 2018.

POINTON, A. & SANGWIN, C. J. (2003) An analysis of undergraduate core material in the light of hand-held computer algebra systems. *International Journal of Mathematical Education in Science and Technology*, **34**, 671–686.

SANGWIN, C. J. (2003) New opportunities for encouraging higher level mathematical learning by creative use of emerging computer aided assessment. *International Journal of Mathematical Education in Science and Technology*, **34**, 813–829.

SANGWIN, C. J. (2013) *Computer aided assessment of mathematics*. Oxford University Press.

SANGWIN, C. J. (2018) High stakes automatic assessments: developing an online linear algebra examination. *Proceedings of 11th Conference on Intelligent Computer Mathematics*. Hagenberg, Austria: Hagenberg, Austria.

SMITH, G., WOOD, L., COUPLAND, M., STEPHENSON, B., CRAWFORD, K. & BALL, G. (1996) Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, **27**, 65–77.

TALLMAN, M. A., CARLSON, M. P., BRESSOUD, D. M., PEARSON, M. & ORG, P. (2016) A Characterization of Calculus I Final Exams in U.S. Colleges and Universities. *International Journal of Research in Undergraduate Mathematics Education*, **2**, 105–133.

WOOD, L. N., SMITH, G. H., PETOCZ, P. & REID, A. (2002) Correlation between Student Performance in Linear Algebra and Categories of a Taxonomy. *Proceedings of the 2nd International Conference on the Teaching of Mathematics at Undergraduate Level*. Hersonissos, Crete: Hersonissos, Crete.

WOOD, L. N. & SMITH, G. H. (2002) Perceptions of difficulty. *Proceedings of the 2nd International Conference on the Teaching of Mathematics at Undergraduate Level* (M. Boezi ed.). Hersonissos, Crete: Hersonissos, Crete.

George Kinnear (G.Kinnear@ed.ac.uk) is a lecturer in the School of Mathematics at the University of Edinburgh. His research interests are in effective uses of technology to support undergraduate mathematics teaching.

Max Bennett, Rachel Binnie, Róisín Bolt and Yinglan Zheng are recent graduates of the School of Mathematics at the University of Edinburgh, where they undertook this research as part of a final-year undergraduate project.

Appendix

FIG. 6. Percentage of marks in each MATH Group in each exam paper, as summarised in Figure 2. There are no further details for DiagTest since it is a single paper.